Notat om statistisk inferens

Larsen, Martin Vinæs

Publication date:
2014

Document version
Peer-review version

Note om statistisk inferens

De store tals lov og den centrale grænseværdisætning

Institut for Statskundskab, KU · Metode 1 · Martin Vinæs Larsen

Efterår, 2013

Denne note introducerer to statistiske principper, der er afgørende for, at vi kan sige noget om vores population ud fra en stikprøve: de store tals lov og den centrale grænseværdisætning. Først diskuteres med udgangspunkt i stikprøvesituationen, nogle simple men tilstrækkelige antagelser, der gør de to principper gældende. Herefter beskrives de store tals lov for henholdsvis stikprøvegennemsnittet og stikprøvevariansen. Endeligt beskrives den centrale grænseværdisætning for stikprøvegennemsnittet, samt andre udvalgte gennemsnit. Noten er udviklet i forbindelse med undervisningen på Metode 1, og er tænkt som et supplement til Agresti og Finlay (2013) kapitel 4.

1 Nogle grundlæggende antagelser

I mange situationer vil man stå over for situationer, hvor man kun kan indsamle informationer om en mindre del af den population af enheder, man egentlig er interesseret i. For at komme omkring dette problem udvælger man en stikprøve bestående af n enheder. Disse enheder kan i princippet være alt muligt: avisartikler, skoler, kandidater til kommunalvalg osv. Af præsentationsmæssige årsager vil denne note imidlertid tage udgangspunkt i et eksempel, hvor man tager en tilfældig stikprøve af individer. De konklusioner som noten kommer frem til, gælder imidlertid for alle tilfældige stikprøver.

Når man først har fået fat i en stikprøve af enheder, så vil man måle disse på en række relevante variable. Lad os, for at holde det simpelt, tage udgangspunkt i én variabel, X. Al vores information om det enkelte individ, i, er nu identificeret ved sin værdi på denne variabel, X_i. Det kunne eksempelvis være respondentens ideologiske orientering på en skala der går fra 1 (venstreorienteret) til 10 (højreorienteret). Med udgangspunkt i sin stikprøve kunne man nu udregne *stikprøvegennemsnittet* som

$\frac{\sum_i^n x_i}{n} = \bar{x}$, ligesom man kunne udregne *stikprøvevariansen* som $\frac{\sum_i^n (x_i - \bar{x})^2}{n} = s^2$. [1] Hvis vi vender tilbage til danskernes ideologiske orientering, så kunne vi altså med udgangspunkt i vores stikprøve af danskere, udregne hvor de gennemsnitligt lå på tipunktsskalaen samt hvor spredte deres svar var (variansen).

Vi er imidlertid ikke synderligt interesserede i $\bar{x}$ eller s^2. De beskriver jo bare vores stikprøve. Vi er i virkeligheden interesseret i det gennemsnit, som findes blandt alle enheder, μ, og den variation der findes blandt alle vores enheder, σ^2. For at kunne sige noget om μ og σ^2 bliver vi imidlertid nødt til at gøre os en central antagelser om, hvordan vores stikprøve er udvalgt.

Den vigtigste antagelse er, at stikprøven er udvalgt tilfældigt. Tilfældig udvælgelse er vigtig, da den sikrer at alle populationens enheder vil have lige stor sandsynlighed for at blive udvalgt til ens stikprøve. Når alle enheder har samme sandsynlighed for at blive udvalgt, vil sandsynligheden for at en udvalgt enhed har et givent udfald, være lig andelen af enheder, der har dette udfald i populationen. Hvis vi vender tilbage til den ideologiske orientering, så vil sandsynligheden for at udvælge en person i stikprøven, der har værdien 1 på skalaen, svare præcist til andelen af alle danskere der har værdien 1. [2]

Tilsammen gør disse to antagelser, at vi kan forstå vores stikprøve som n *uafhængige træk* af den variabel vi er interesseret i, X, hvor de enkelte udfald, X_i, er trukket fra en sandsynlighedsfordeling med populationens gennemsnit (μ) og varians (σ^2).

2 De store tals lov

Hvis man udvælger enheder tilfældigt fra ens population, når man trækker sin stikprøve, kan det altså forståes som at lave uafhængige træk fra en sandsynlighedsfordeling. Det er en produktiv måde at forstå stikprøveudvælgelse på, fordi vi hermed kan bruge nogle teoremer fra sandsynlighedsregningen til at beskrive forholdet mellem vores stikprøve og vores population. Det første teorem vi skal se på er *det store tals lov*, der siger følgende om forholdet mellem vores stikprøvegennemsnit ($\bar{x}$) og gennemsnittet i vores population (μ):

[1] Hvis formlerne for hhv. gennemsnit og varians ikke virker bekendte, så konsulter kapitel 3 i Agresti og Finaly (2013). Bemærk derudover at nævneren i formlen for stikprøvevariansen er n og ikke n-1. Det skyldes, at vi i denne note udelukkende beskæftiger os med asymptotiske resultater (se afsnit 3.2), hvor n går mod uendelig, og når n går mod uendelig vil $n \approx n - 1$.

[2] Denne skal teknisk set suppleres med en antagelse om at vores stikprøve er udvalgt på baggrund af en populationsfordeling med et bestemt gennemsnit, μ, og en bestemt varians, σ^2. Denne antagelse vil sjældent være brudt i praksis, og har derfor en mere teoretisk karakter. Den etablerer blot, at der er et gennemsnit og en varians "derude", som vi gerne vil sige noget om.

$$\bar{x} \to \mu \tag{1}$$

når $n \to \infty$

$\bar{x}$ vil altså nærme sig populationens gennemsnit, μ, efterhånden som vores stikprøve bliver større - altså når n går mod uendelig. Et interessant og intuitivt resultat.

Intuitivt fordi mere information om gennemsnittet i vores population (større n), medfører at vores stikprøvegennemsnittet kommer tættere på populationens gennemsnit.

Et interessant resultat, fordi vi nu ved, at jo flere vi udvælger til vores stikrprøve, desto tættere kommer vi på populationens gennemsnit. Hvis vi igen vender tilbage til danskernes højre-venstre placering, så kommer vi altså tættere og tættere på danskernes gennemsnitlige ideologiske orientering, når vi gør vores stikprøve større.

Det store tals lov gælder for en lang række stikprøvemål. Det kræver blot, at vi kan omkskrive strikprøvemålet til et gennemsnit. Lad os tage et eksempel, der kommer til at have en naturlig interesse. Gennemsnittet af de kvadrerede afvigelser fra gennemsnittet, $\frac{\sum_i^n (x_i - \bar{x})^2}{n} = s^2$.

Vi kalder dette gennemsnit for s^2, da formlen er den samme som formlen for variansen i stikprøven (jf. afsnit 1). Ligesom overfor kan vi anvende de store tals lov og komme frem til følgende:

$$s^2 \to \sigma^2 \tag{2}$$

når $n \to \infty$

Ligesom at stikprøvegennemsnittet vil nærme sig populationens gennemsnit når n går mod uendelig, så vil stikprøvevariansen også nærme sig populationens varians.

3 Den centrale grænseværdisætning

Selvom det kan være svært at få armene ned oven på de store tals lov, så står vi stadig tilbage med et problem. Vi vil aldrig kunne udtrække uendelig observationer, så vores stikprøvegennemsnit vil aldrig blive præcis lig populationens gennemsnit. Så selvom vi ved, at vores stikprøvegennemsnit vil komme tættere på, hvis vi indsamler mere data om danskers ideologi, så ved vi ikke noget om, *hvor* langt vi er fra populationens gennemsnit. For at kunne løse dette problem skal vi sige noget om hvilken sandsynlighedsfordeling, som stikprøvegennemsnittet , $\bar{x}$, følger. Men hvad er nu det? Var det ikke kun

populationen, der havde en bestemt sandsynlighedsfordeling? Nej, faktisk har stikprøvegennemsnittet også en sandsynlighedsfordeling.

Det kan umiddelbart virke lidt kontraintuitivt at tænke stikprøvegennemsnittet som noget der kan variere. Typisk vil man jo kun have én stikprøve, og således også kun ét stikprøvegennemsnit. Man kan imidlertid forstå stikprøvegennemsnittets sandsynlighedsfordeling, som den fordeling af gennemsnit, du ville se, hvis du blev ved med at udvælge et *nyt* tilfældigt udsnit af dine respondenter og registrere stikprøvegennemsnittet. Ligesom trækkene af X følger en sandsynlighedsfordeling, vil disse stikprøvegennemsnit ($\bar{x}$) også følge en sandsynlighedsfordeling. Hvor populationens sandsynlighedsfordeling beskriver hvor sandsynligt det er, at få et givet udfald på X, så beskriver stikprøvegennemsnittets sandsynlighedsfordeling, hvor sandsynligt det er at få et givent stikprøvegennemsnittet. Hvad ved vi om stikprøvegennemsnittets sandsynlighedsfordeling?

Fra de store tals lov ved vi at den ændrer sig når n bliver større. Vi ved nemlig at dit stikprøvegennemsnit vil komme tættere på den rigtige værdi jo flere observationer du får - altså vil sandsynlighedsfordelingen for stikprøvegennemsnittet variere mindre og mindre efterhånden som n stiger. Når du indsamler uendelig observationer vil fordelingen kollapse helt, og du vil få det sande gennemsnit. Men hvad så når n ikke er uendelig? Hvordan fordeler afvigelser fra det sande gennemsnit sig så? Det fortæller et andet teorem os: *den centrale grænseværdisætning*. Dette teorem beskriver hvordan afstanden mellem stikprøvegennemsnittet og dens sande værdi ($\bar{x} - \mu$) er fordelt:

$$(\bar{x} - \mu) \sim N(0; \sigma^2/n) \tag{3}$$

$$\text{når } n \to \infty$$

Grundlæggende siger den centrale grænseværdisætning altså, at efterhånden som n går mod uendelig, så vil forskellen mellem stikprøvegennemsnittet og populationens gennemsnit ($\bar{x} - \mu$) følge en normalfordeling, der gennemsnitligt vil være nul og have en varians, der er lig populationens varians divideret med n. Hvad betyder det?

For det første, og som vi vidste fra de store tals lov, så vil den gennemsnitlige forskel mellem stikprøvegennemsnittet og populationens gennemsnit være lig nul. Gennemsnitligt set vil $\bar{x}$ altså være lig μ.

For det andet, og mere interessant, kan vi nu også sige noget om variansen af afvigelsen mellem stikprøvegennemsnittet og populationens gennemsnit - altså hvor langt $\bar{x}$ typisk vil ligge fra μ. Konkret vil variansen af afvigelserne være lig variansen i populationen divideret med n.

Endelig ved vi at afvigelserne fra populationsgennemsnittet vil være normalfordelte. Når en variabel er normalfordelt med et bestemt gennemsnit og en bestemt varians, så betyder det at vi kan knytte bestemte sandsynligheder til, at vi får bestemte udfald på denne variabel. I det her tilfælde er vores variabel de forskellige afvigelser fra populationsgennemsnittet vi kunne have fået, såfremt vi havde spurgt et andet tilfældigt udsnit, og vores udfald er den afvigelse, vi rent faktisk har fået. Vi kan altså knytte en bestemt sandsynlighed til, at vi får en bestemt afvigelse fra populationsgennemsnittet. Og her når vi så til sagens kerne - hvorfor den centrale grænseværdisætning er så *central* - den kan, for en given stikprøvestørrelse, fortælle os hvor sandsynligt det er, at vi lander et bestemt stykke fra populationsgennemsnittet. Vi kan altså, hvis vi spørger 500 danskere, udregne, hvor sandsynligt det er, at vores stikprøvegennemsnit er landet et (eller to) skalapunkter fra det sande gennemsnit.

3.1 Nogle vigtige omskrivninger af den centrale grænseværdisætning

Det er sådan set alt man behøver vide om den centrale grænseværdisætning. Det er imidlertid relevant at gennemgå et par typiske omksrivninger af sætningen, der tillader os at bruge den mere effektivt, og eksplicitere intuitionen bag teoremet.

Vi starter med at se på variansen af afvigelserne fra populationsgennemsnittet, σ^2/n. Den fortæller grundlæggende noget om hvor præcise vores gæt (dvs. stikprøvegennemsnittet) på populationens gennemsnit er. Konkret bliver vi mindre præcise, når populationsvariansen stiger σ^2 . Når det er mere sandsynligt at et givet X_i ligger langt fra μ, er X_i mindre informativ i forhold til μ. Derudover så falder variansen når n stiger, hvilket vi også ville forvente på baggrund af de store tals lov. Jo flere observationer vi får, jo tættere vil vi ligge på det rigtige gennemsnit.

Et problem ved variansudtrykket, σ^2/n, er at vi generelt ikke kender[3] σ^2. Vi ved imidlertid fra de store tals lov at $s^2 \to \sigma^2$ når $n \to \infty$. Derfor kan vi udskifte den ukendte varians σ^2 med den kendte stikprøvevarians s^2, når n går mod uendelig:

$$(\bar{x} - \mu) \sim N(0; \hat{s^2}/N) \tag{4}$$

$$\text{når } n \to \infty$$

En anden omksrivning går ud på at standardisere afvigelserne fra populationsgennemsnittet. Dette gøres ved at dividere med fordelingens standardafvigelse, der er lig kvadratroden af variansen (σ). Så ser den centrale grænseværdisætning ud som følger:

[3] Præcis ligesom vi ikke kender populationsgennemsnittet kender vi heller ikke populationsvariansen.

$$\frac{(\bar{x} - \mu)}{\hat{\sigma}/\sqrt{N}} \sim N(0; 1) \tag{5}$$

når $n \to \infty$

Hvor udtrykket til venstre ofte blot betegnes z, og kan forstås som antallet af standardafvigelser som stikprøven lander fra stikprøvegennemsnittet.

$$z \sim N(0; 1) \tag{6}$$

når $n \to \infty$

Styrken ved denne omskrivning er at z-værdierne nu vil være standardnormalfordelte. Det vil sige, at afvigelserne er normalfordelte med gennemsnit 0 og varians 1. Dette er en særlig simpel normalfordeling, som let kan bruges til at udregne sandsynligheden for forskellige z-værdierne. Man kan eksempelvis finde sandsynligheden for en given z-værdi i Agresti og Finlay tabel A (2013, 592)

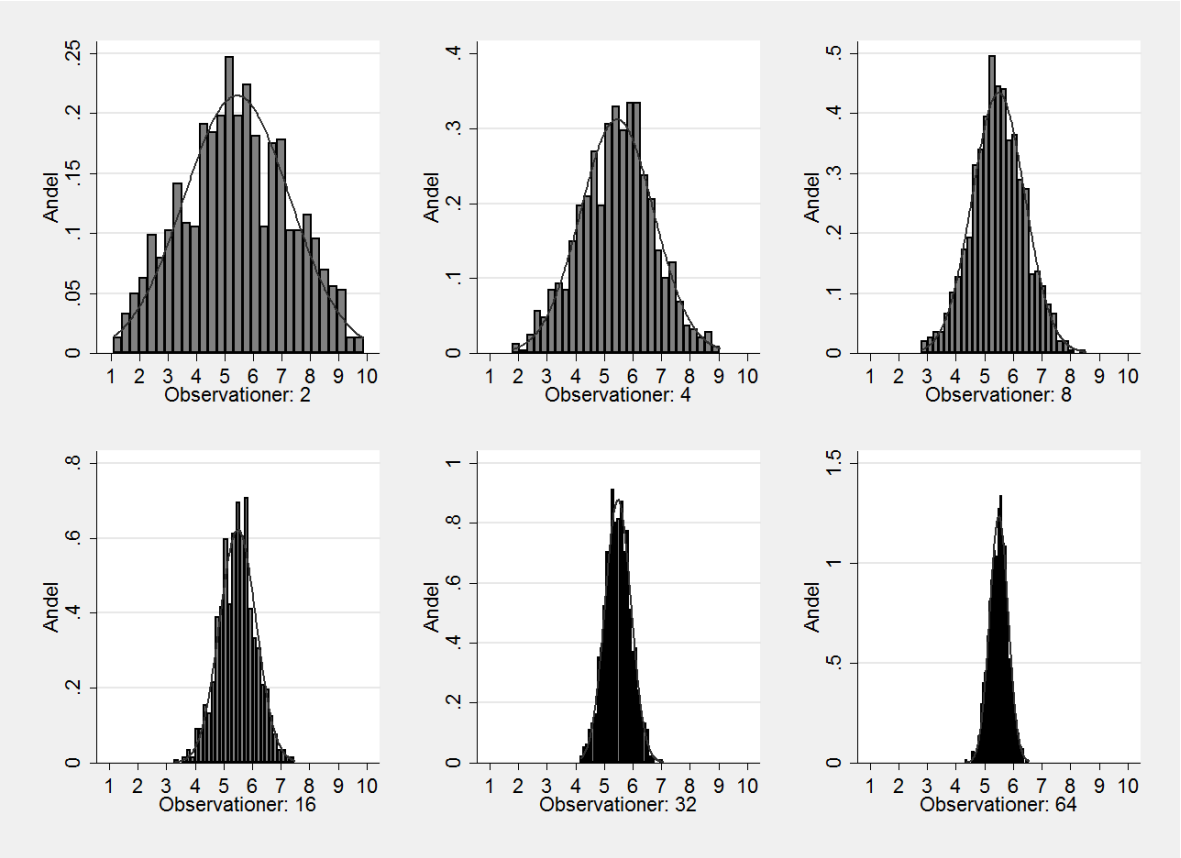
3.2 Hvor stor skal n være?

Alle resultater ovenfor har været det man kalder asymptotiske. Det vil sige, at de kun er rigtige når $n \to \infty$, men hvad betyder det rent faktisk når teoremerne anvendes? Ikke så meget. Det viser sig nemlig, at n ikke skal være så langt mod uendeligt, før at den centrale grænseværdisætning er ret god til at beskrive stikprøvegennemsnittets afvigelser fra gennemsnittet. Agresti og Finlay (2013) anbefaler 30 observationer, men det faktiske antal varierer afhængigt af populationens fordeling. Et eksempel kan findes i figur 1. Her er trukket 1000 stikprøver af hhv. 2, 4, 8, 16, 32 og 64 observationer fra en population af individer, der er helt ligeligt fordelt på tværs af en højre-venstre skala fra 1-10, hvorfor gennemsnitet er på 5,5. Figuren viser, hvordan stikprøvegennemsnittet fordeler sig på tværs af de 1000 stikprøver. Som det kan ses, begynder normalfordeling allerede at være en fornuftig approksimation ved omkring 4-8 observationer. Bemærk desuden, hvordan stikprøvegennemsnittene ligger tættere og tættere på det faktiske gennemsnit, efterhånden som stikprøven bliver større.

3.3 Den centrale grænseværdisætning og dummyvariable

Hvis man ønsker at estimerer sandsynligheden π for et udfald (a) kan vi ligeledes bruge den centrale grænseværdisætning. Det kan eksempelvis være sandsynligheden for at en tilfældig dansker er Socialdemokrat. Vi definerer i dette tilfælde en variabel X med kun to udfald {0,1}, hvor det udfald vi er

Figur 1: Hvor stor skal n være? En illustration



interesserede i gives værdien 1, og alle andre udfald værdien 0. En sådan variabel kaldes en dikotom, binær eller dummy-variabel. Lad os nu forestille os, at vi har udvalgt en tilfældig stikprøve af danskere. Hvis du vil udregne sandsynligheden p for at du trak en socialdemokrat fra din stikprøve, ville du blot skulle tage antallet af det udfald du var interesseret i (Socialdemokrater) og dividerer med samtlige udfald (din stikprøvestørrelse), matematisk: $p = \sum_i^n X_i/n$.

Dette er præcis udtrykket for vores stikprøvegennemsnittet. Derfor gælder det store tals lov:

$$p \to \pi \tag{7}$$

når $n \to \infty$

Det samme gør den centrale grænseværdisætning:

$$(p - \pi) \sim N(0; s^2) \tag{8}$$

når $n \to \infty$

Stikprøvevariansen kan i dette tilfælde også udtrykkes endnu meres simpelt, nemlig som $p(1-p)$, hvorfor følgende udtryk kan bruges:

$$(p - \pi) \sim N(0; p(1-p)) \tag{9}$$

når $n \to \infty$